Identifying Transcription Regulatory Elements in the Human and Mouse Genomes Using Tissue-specific Gene Expression Profiles

Amitava Karmaker[1], Kihoon Yoon[1], Mark Doderer[1], Russell Kruzelock[2], Stephen Kwek[1*]

[1] Department of Computer Science, University of Texas at San Antonio,
San Antonio, Texas 78249, USA
[2] The Institute for Drug Development, Cancer Therapy and Research Center,
San Antonio, Texas 78245, USA

Summary

Revealing the complex interaction between *trans*- and *cis*-regulatory elements and identifying these potential binding sites are fundamental problems in understanding gene expression. The progresses in ChIP-chip technology facilitate identifying DNA sequences that are recognized by a specific transcription factor. However, protein-DNA binding is a necessary, but not sufficient, condition for transcription regulation. We need to demonstrate that their gene expression levels are correlated to further confirm regulatory relationship. Here, instead of using a linear correlation coefficient, we used a non-linear function that seems to better capture possible regulatory relationships. By analyzing tissue-specific gene expression profiles of human and mouse, we delineate a list of pairs of transcription factor and gene with highly correlated expression levels, which may have regulatory relationships. Using two closely-related species (human and mouse), we perform comparative genome analysis to cross-validate the quality of our prediction. Our findings are confirmed by matching publicly available TFBS databases (like TRANFAC and ConSite) and by reviewing biological literature. For example, according to our analysis, 80% and 85.71% of the targets genes associated with *E2F5* and *RELB* transcription factors have the corresponding known binding sites. We also substantiated our results on some oncogenes with the biomedical literature. Moreover, we performed further analysis on them and found that *BCR* and *DEK* may be regulated by some common transcription factors. Similar results for *BTG1, FCGR2B* and *LCK* genes were also reported.

1 Introduction

The completion of the Human Genome Project (HGP) [1-3] signifies not the end of a journey but rather the beginning of an exciting expedition in revealing the secret of the human genome. To be truly benefited from the HGP, we still need to decipher the biological connotation of the sequences; otherwise, these are just some long strings of meaningless letters. The natural next step is to identify the functional elements in the human genome and understand how the genes regulate and interact with each other. While a large number of human genes have been identified, their regulatory mechanism remains mostly unknown even at the transcriptional level[4].

The main mechanism of transcriptional control is to bind transcription factors (TF) to *cis*-elements (TFBS a.k.a. transcription factor binding sites) either upstream or downstream of the regulated gene, scattered all along thousands of base pairs in both intergenic and intragenic regions. In doing so, it either enhances or suppresses gene transcription. To some extent,

[*] Corresponding author, kwek@cs.utsa.edu

trans-elements can be viewed as “keys” needed to unlock the *cis*-elements which act as “locks”. To comprehend gene transcription mechanism, it is not sufficient to know which keys (*trans*-elements) are needed to lock/unlock a specific gene, but we also need to identify their corresponding locks (*cis*-elements). Moreover, identification of binding sites serves as a form of validation on the putative *trans*-elements. If we are able to identify a putative *cis*-element, say a conserved sequence element, which is over-represented in the genes that we believe to be regulated by a particular transcription factor then it is more plausible that the TF acts as a *trans*-element.

In silico discovery [5] of binding sites is quite effective for prokaryotes, like ***Escherichia coli*** [6, 7], where genomes are more compact with many genes being regulated by a single operator that is relatively easy to locate. Similar successes have been reported for simple unicellular eukaryotes, like ***Saccharomyces cerevisiae*** [8-10]. The main approach for finding regulatory elements of such simple organisms is to search for overrepresented motifs modeled by known background profiles, such as position weighted matrices (PWMs) [11-17], position specific score matrices (PSSMs) [11, 18, 19], while some use clustering to demarcate cis-regulatory modules[20, 21]. For higher multi-cellular eukaryotes, model-based approaches [4, 22-24] that discover patterns among co-expressed genes with respect to regulating transcription factors, have been proposed. The idea behind these techniques involves the proximity of common cis-regulatory modules among the co-expressed genes. Among other common model-based (a.k.a. machine learning) techniques, artificial neural networks [25], greedy algorithm [12], Gibbs Sampling [26], Markov chains [27, 28], Expectation Maximization (EM) algorithm [29] are widely used for eukaryotes. However, it has been reported that these model-prediction techniques are susceptible to high false positive prediction rate and majority of predicted TFBS generated with predictive models (*in silico*) have no functional role *in vivo* [14].

To overcome this potential shortcoming, quite a number of algorithms have been presented motivated by “phylogenetic footprinting” [30-33], which is based on searching for similarity in sequences due to selective pressure during evolution,. In fact, phylogenetic footprinting complements the computational approaches, as sequence conservation across lineage reveals segments in genes that might delineate common biological functions. So, to identify regulatory regions, orthologous genes at the sequence level are compared. The underlying hypothesis that inspires phylogeny as a powerful scheme is that sequences related to vital biological functions will be retained under evolutionary selective pressure [34, 35]. It is assumed that the orthologous genes are accountable to the same regulatory mechanisms in different species. Again, phylogenetic footprinting algorithms can detect putative binding sites if it meets two criteria: (1) sequences from organisms with adequate evolutionary distance share the same conserved regions, and (2) TFBS are over-represented in the proximity. Thus novel methods using cross-species genome comparison can significantly improve the overall specificity of predictions [36, 37].

Jin et al. [38, 39] analyzed conserved human-mouse orthologous gene pairs to find core promoter elements and Bussemaker et al. [23] addressed the issue of detecting regulatory elements using correlation of expressions. A recent paper by Kim et al. [40] dealt with predicting transcriptional regulatory elements of human promoters using gene expression and promoter analysis data, which compare two pools of genes using z-scores. Our aim here is to find both *trans*- and *cis*-elements in the human genome. Further, to ensure better quality of our analysis, we also used mouse tissue-specific gene expression profile and the observation that functional elements tend to be conserved between mouse and human.

As a model organism to compare with, the choice of mouse is obvious [31]. Human and mouse have about 25,000 genes each, of which about eighty percent coincide in DNA sequence abstraction [41]. Along with progress in sequencing and annotations of mouse genome [42, 43], genome-scale interspecies comparison techniques have evolved [44], and geneticists have been amazed by the incredible resemblance between human and mouse genomes. In numbers, almost 40% of total nucleotides show exact matches in global alignments [45], and about 93.2% of identical nucleotides has been observed in conserved regions [46]. Moreover, the mouse-model is extensively used to study the molecular foundation and therapy of certain diseases, as functionally related genes are preferentially retained in conserved regions [47].

Based on the flavor of the both strategies of phylogenetic footprinting and co-regulated gene expression profiles, in this paper we propose a systematic approach to identify putative transcriptional regulatory elements in both human and mouse genomes. Our algorithm delineates potential TF-gene pairs by analyzing tissue-specific gene expression data of both human and mouse. As a means of validation, we restrict our attention to orthologous TFs and genes that share a common gene symbol between the two species, and validate our findings by comparing against a widely used TFBS profile registry and reviewed literature.

To investigate the efficacy of our technique and utility of our findings, we computed the correlation among TFs and well-recognized oncogenes to analyze their regulation pattern collectively. The objective here is to identify a small collection of common TFs that regulate a group of oncogenes. For instance, we observed that leukemia proto-oncogenes like *BCR* and *DEK* may be regulated sharing some common transcription factors. We also have similar results for the genes *BTG1, FCGR2B* and *LCK*. These results together with current biomedical literatures serve to corroborate our findings. Additionally, we also clustered the oncogenes based on their correlation with TFs and the results we report here support our earlier analysis.

2 Methods

2.1 Data preprocessing

As mentioned earlier, we analyzed expression data for both human and mouse to shortlist co-expressed genes with respect to TF. We collected microarray data of normal human tissues[48], which provide us with 26,260 unique genes from 35 different organs. In total, the data set consists of 115 tissue specimens. The tissue samples were obtained from esophagus, lungs, cervix, fallopian tube, breast, ovary, uterus, seminal vesicles, prostate, salivary glands, thyroids, parathyroid, testes, brain, adrenal, heart, stomach, colon, bowel, kidney, liver, lymph, tonsil, spleen, buffy coat, thymus, vagina, skeletal muscle, fallopian tube, uterus and gall bladder. For each experimental tissue sample, Cy5- and Cy3- labeled samples were co-hybridized to a cDNA microarray containing 39,711 human cDNA's, representing 26,260 different gene. Expression ratios were globally normalized by mean-centering each gene across all arrays.

Raw microarray expression data for mouse was obtained from [49]. The dataset contains expressions of 41,699 known and predicted mRNAs in 55 tissues. The samples include adrenal, aorta, bladder, bone marrow, brain, brown fat, calvaria cerebellum, colon, cortex, embryo, epididymus, eye, femur, heart, hindbrain, kidney, knee, large intestine, liver, lung, lymph node, mammary gland, mandible, midbrain, olfactory bulb, ovary, pancreas, placenta, prostate, salivary, skeletal muscle, skin, small intestine, snout, spinal cord, spleen, stomach, striatum, teeth, testis, thymus, thyroid, tongue, tongue surface, trachea, trigeminus, uterus etc. As the expression annotations were given for NCBI XM sequences, we had to map those to official NCBI gene[50] annotation. Although the NCBI repository is being updated and adapted, the RefSeq[51] accession number (e.g. XM_213390) may remain unchanged over time. Therefore, direct one-to-one correspondence of gene id and accession number does not produce correct gene list. To circumvent this problem, we retrieved the GI (gi) numbers (GenInfo Identifier), which are exclusively assigned to the sequence records. This number also provides a unique identifier to the sequences, and it is not reused across different versions of accession numbers.

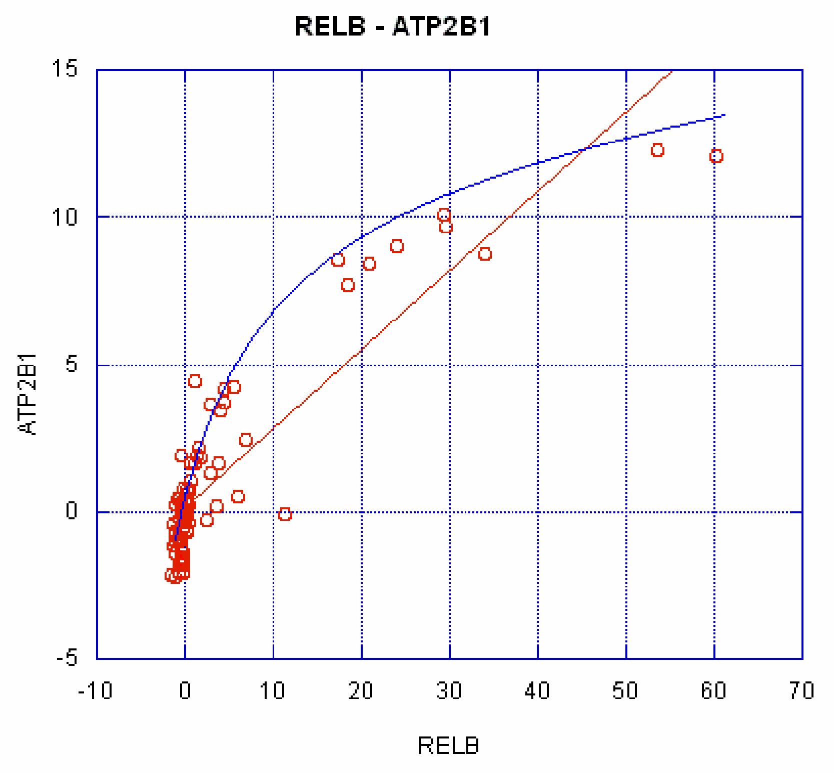


Figure 1: Illustration for TF-gene regulation pattern. Here, the expression data are plotted for TF (RELB) along x-axis, and for gene (ATP2B1) along y-axis. Linear regression (straight line) is unable to capture majority of points, while the non-linear function curve seems to cover most of them.

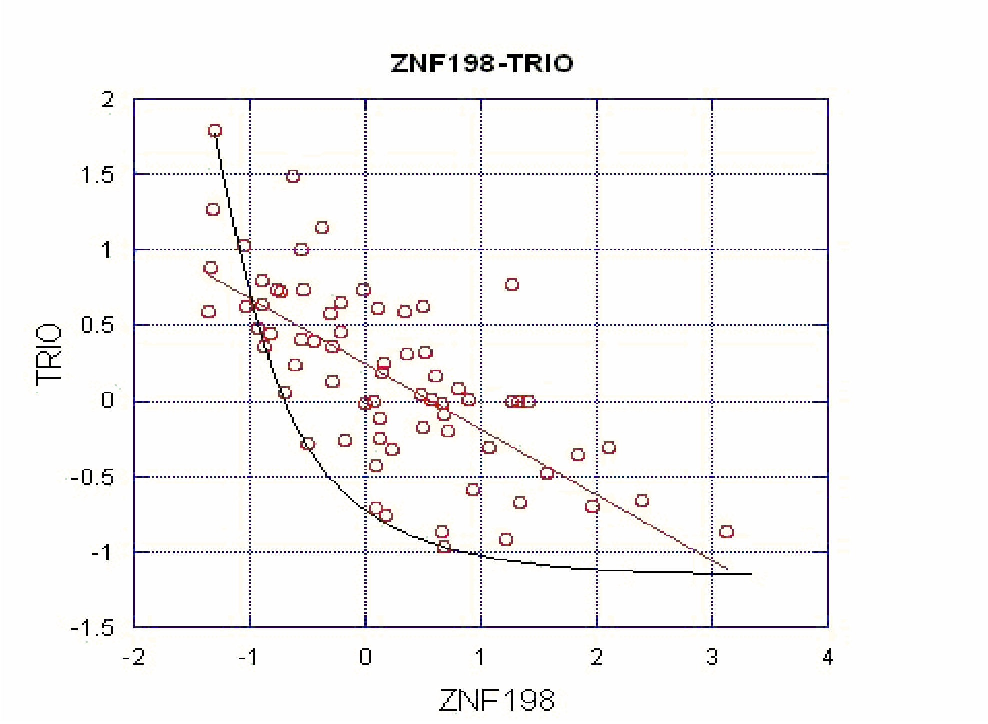


Figure 2: Illustration for negatively correlated TF-gene regulation pattern.

2.2 Calculation of correlation coefficient

If a transcription factor does regulate a gene, according to reported results [12, 24] in the literature, it is expected that they are linearly correlated. However, we observed that very often there seems to be a saturation point where the effect on the expression level of the gene diminishes as the level of transcription factor continues to increase and may reach a plateau or even decrease in some cases (see Figure 1). Again, Figure 2 shows the regulation pattern for inversely correlated TF and gene. In these plots, positive numbers represent the up-regulation of genes (or TF), while negative magnitudes for numbers stand for down-regulation of genes. Thus, instead of using simple linear correlation, we measure the correlation using Equation 1 as our regression curve.

$$y' = ye^{\alpha y}, \quad \text{Where } \alpha \text{ is exponential constant} \qquad \textbf{(1)}$$

In this equation, y' is dependent on the independent variable y, which is the original expression level, and it is multiplied by some exponential constant to generate new values (y'). This equation is used commonly in physics to measure half-life to measure radioactive decay. The value of parameter α was set to 0.0, 0.25 and 0.5 in different experimental settings. This correlation coefficient is more general than simple linear correlation coefficient in that by setting α = 0.0, we end up with the simple linear correlation coefficient.

We calculated Pearson's Correlation Coefficient, of all pairs of gene and TF for both human and mouse datasets using the pre-calculated expression numbers. The values of Pearson's Correlation Coefficient range from -1 to +1. Any value in positive scale indicates increasing correlationship, with +1 being perfectly linear correlated and negative values denote the case of a negative correlationship. Any value in between in all other cases represents the degree of dependency between the variables (i.e. gene and TF pair). The correlation coefficients indicate how tightly genes are up-regulated and down-regulated with respect to transcription factors.

2.3 Finding cis-regulatory elements

To determine the (putative) *cis*-regulatory elements, we take advantage of the fact that our analysis is from tissue-specific microarray of two mammalian species, human and mouse. Thus, we can identify orthologous genes between the two species that we speculate to be associated with a common TF and analyze their promoter regions using sequence alignment to determine possible binding site for the TF [34]. For the human genome, we associate a TF as a possible enhancing or repressing regulatory element for a gene if the correlation coefficient is greater than 0.3 or less than -0.3, respectively. For the mouse genome, we have a more stringent threshold as the correlation coefficients are fairly high. The thresholds for positive and negative correlations are set at 0.4 and -0.4, respectively.

Table 1: List of top ten highly correlated (both positively and negatively) human genes with corresponding transcription factors (here, α = 0.25).

Transcription Factor	Gene	Correlation Coefficient
SCAND1	*ASGR1*	0.9460
NOTCH2	*INSIG1*	0.9403
ID1	*ITGB1*	0.9326
TCEB3	*CTGF*	0.9308
STAT4	*S100A8*	0.9279
ZNF134	*DSTN*	0.9270

RELB	*ATP2B1*	0.9268
RELB	*PSMB9*	0.9190
RBPSUH	*SSR4*	0.9173
SCAND1	*F2*	0.9151
ZNF198	*TRIO*	-0.6928
PPARBP	*CENPB*	-0.6882
GABPA	*C22orf5*	-0.6831
HCLS1	*C4A*	-0.6803
ZNF228	*EHD1*	-0.6800
TEAD3	*ARFGEF2*	-0.6705
PPARBP	*EPN1*	-0.6695
HCLS1	*C4A*	-0.6683
PPARBP	*BCR*	-0.6650
GABPA	*NCOR2*	-0.6645

Transcriptional regulatory elements are found either upstream or downstream of genes, scattered all along thousands of bps in both intergenic and intragenic regions. However, most TFBS predictors tend to focus in the proximal promoter region [52] because the difficulty of TFBS prediction tends to increase with the size of the region of interest. Besides, increasing the region of interest upstream of the transcription start site to more than a few thousand base pairs increases the chances of falsely identifying common repeat elements. Thus, we focus on the core promoter regions from 500 bps upstream to 200 bps downstream (-500 to +200, total 700 bps).

To ensure that our putative TF binding sites are of high quality, we validated them with TRANSFAC database [24, 53, 54]. TRANSFAC contains the largest repository for experimentally derived TFBS. We also performed further validations of our putative sites using P-Match [55]-public and ConSite [56], which combines pattern matching and weight matrix approaches thus providing higher accuracy of recognition than each of the methods alone. To reduce false-positive validation using P-match, we chose “high quality vertebrate matrices only” as our default option. We obtained the report for all pre-selected genes, setting cut-off selection for matrices to minimize (1) false-positive, (2) false-negative, and (3) the sum of both error rates. Moreover, ConSite [56] is a user-friendly, web-based tool for finding cis-regulatory elements in genomic sequences using high-quality transcription factor models and cross-species comparison filtering.

2.4 Clustering the oncogenes

In order to explore the collective interactions among TF and oncogenes, we clustered the oncogenes with respect to regulating TFs. We constructed an interaction matrix, M, where each entry is either one (1) or zero (0) individually based on the correlation coefficients of the respective TF-gene pairs. For each possible pair of genes, if the number of one’s along the rows exceeds a cut-off, we assign them in the same cluster. These pairs of genes will then be used as ‘seeds’ to grow the cluster in a greedy fashion, where new gene is added to the cluster if its putative TFBS are found to contain many of the putative TFBS of the genes already in the cluster. In each cluster, the correlations between TF and oncogenes are analyzed as a group to capture their differential regulation patterns.

3 Results and Discussion

3.1 Correlated genes and transcription factors

Using Pearson's Correlation Coefficient (see Methods section), we constructed a list of TF-gene pairs, either correlated or inversely correlated to each other. Table 1 and Table 2 show the top ten most positively and negatively correlated genes and transcription factors for human (α = 0.25) and mouse (α = 0.5) in order. The complete list of correlated pairs can be obtained from our website [57].

We find that many of the TF-gene pairs fit the non-linear function better, especially for the human genome. Using a threshold of +0.8 on the correlation coefficients, we find that the non-linear function seems to better depict the possible relationship between the TFs and genes, especially for the human genome. In our dataset for human, we have 1,114,432 possible TF-gene pairs. Of them, the number of correlated pairs are 171 (topmost 0.015%), 372 (0.033%) and 253 (0.023%) for α = 0.0 (linear), α = 0.25 and α = 0.50 respectively. Further, we observed that the correlation coefficients of most pairs are greater when α = 0.25 as opposed to α = 0.0. However, the effect on the mouse genome is not as dramatic, which is partly, we believe, due to the already higher linear correlation coefficient. Out of 4,52,076 possible TF-gene pairs, we found 2,448 (topmost 0.541%), 3,150 (0.70%) and 3,472 (0.77%) correlated TF-gene pairs for corresponding α = 0.0 (linear), α = 0.25 and α = 0.50 above +0.8 correlation coefficient. Therefore, our results concentrate on the correlated TF-gene pairs for human at α = 0.25, and for mouse α = 0.5. One of the novelties of our approach is that the technique can be scaled up by associating more genomes in the lineage, provided *cis*-elements are spatially conserved.

In the results for human expression data, we found that *SCAND1* (SCAN domain containing 1) and *RELB* (v-rel reticuloendotheliosis viral oncogene homolog B) appear more often than others. *SCAND1* is positively correlated with both *ASGR1* (Asialoglycoprotein receptor 1) and *F2* (Coagulation factor II (thrombin)), while *RELB* shows the same pattern with *ATP2B1* (ATPase, Ca++ transporting, plasma membrane 1) and *PSMB9* (Proteasome subunit, beta type, 9). In fact, the interactions between *RELB* and *PSMB9* are reported in the literature [58]. As shown in Figure 1, the regulation pattern among TFs and genes tends to better fit the decay function curve than simple linear line, which characterizes our assumption regarding TF-gene interaction model. In fact, we observed that the correlation coefficient of most TF-gene pairs tend to be higher for the non-linear function regressor than linear regressor.

Table 2: List of top ten highly correlated (both positively and negatively) mouse genes with corresponding transcription factors (here, α = 0.5).

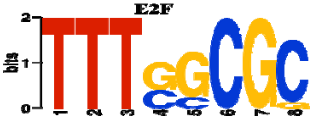
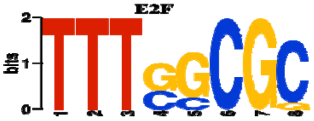
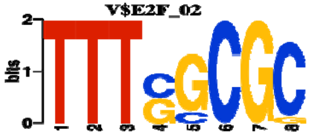
Transcription Factor	Gene	Correlation Coefficient
LEF1	*EPM2A*	0.9946
HOXB1	*1700014N06RIK*	0.9874
FOXE1	*DFNA5H*	0.9825
RFX1	*MESP1*	0.9816
FOXE1	*2610207F23RIK*	0.9799
FOXE1	*5730403I07RIK*	0.9778
CREB1	*2810025M15RIK*	0.9750
E4F1	*4930424G05RIK*	0.9733
ZIC1	*GTF2E2*	0.9720
ACE	*1700027M21RIK*	0.9718
LEF1	*REN*	-0.9545
MBP	*5930416I19RIK*	-0.9501

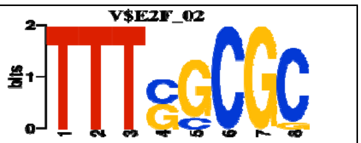
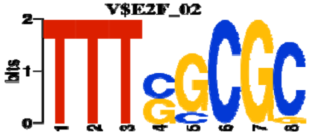
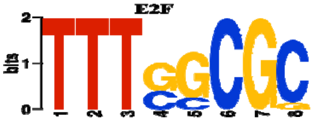
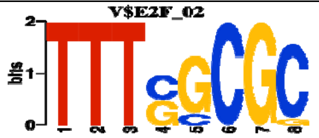
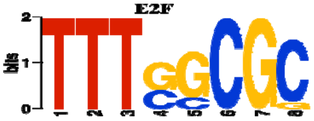
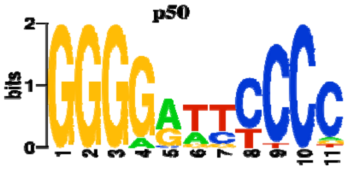
FOXM1	*LOC238771*	-0.9494
FOXD2	*8030423J24RIK*	-0.9442
AP1M1	*P2RY2*	-0.9386
MBP	*2700099C18RIK*	-0.9354
ZIC2	*CLDN4*	-0.9345
SIN3B	*KCNJ11*	-0.9339
HOXB1	*9030420J04RIK*	-0.9297
PITX1	*MCF2*	-0.9167

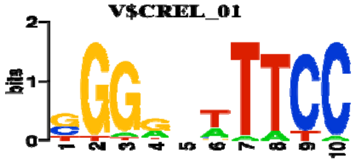
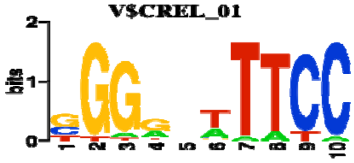
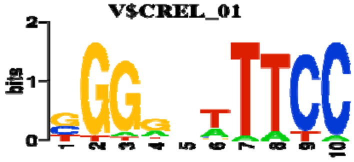
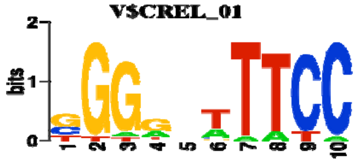
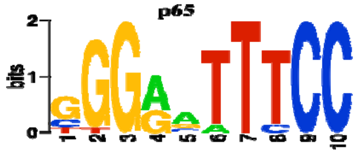
However, in the cases of anti-correlations, the coefficient numbers of TF-gene pairs are relatively low, which is possibly due to intrinsic experimental intricacy in measuring negativity in the gene expressions. Among the top short-listed pairs, *PPARBP* (PPAR binding protein), *GABPA* (GA binding protein transcription factor, alpha subunit 60kDa) and *HCLS1* (Hematopoietic cell-specific Lyn substrate 1) showed up more frequently with corresponding down-regulated genes.

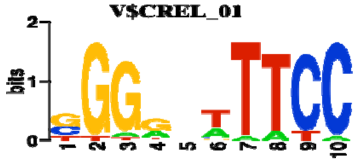
Our analysis of mouse expression data revealed that there are surprisingly strong correlations among genes in both directions, with 3,684 pairs having coefficient greater than 0.8 or below -0.8 as opposed to 372 in human. We believe that this is probably an artifact of the data. *FOXE1* (Forkhead box E1) showed up most often in the positively correlated pairs, while *MBP* (Myelin basic protein) is the most frequently occurring in the negative list. Some TFs seemed to enhance the expression level of some genes while suppressing others. For example, *LEF1* (Lymphoid enhancer-binding factor 1) and *HOXB1* (Homeo box B1) are included in both positive and negative list.

Table 3: The list of identified binding sites for four TFs. The consensus sequence logo plots for available PWM's in TRANSFAC are shown. Results were validated using both TRANSFAC and ConSite.

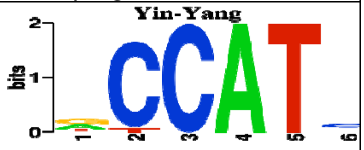
For TF *E2F5*				
Genes	**Factor Name**	**Position in Human sequence (strand)**	**Position in Mouse sequence (strand)**	**Consensus sequence**
EZH2	**TRANSFAC (P-Match):** E2F **ConSite :** E2F	533 (+)	494 (+)	 **TTTGGcgc**
RPA2	**TRANSFAC (P-Match):** E2F **ConSite :** E2F	331 (+)	158 (+)	 **TTTGGcgc**
TYMS	**TRANSFAC (P-Match):** E2F **ConSite :** E2F	464 (-)	468 (-)	 **gcgGGAAA**
CRTAP	**No Hits**			

SUPT16H	**TRANSFAC (P-Match):** E2F **ConSite :** E2F	559 (-)	3 (-)	 **ccgCGAAA**
NMI	**TRANSFAC (P-Match):** E2F **ConSite :** E2F	553 (+)	626 (+)	 **TTTCGcgg**
GART	**TRANSFAC (P-Match):** E2F **ConSite :** E2F	249 (+)	16 (+) Seq found: **TTTGGTCA**	 **TTTGGcgc**
MBD4	**No Hits**			
DDX6	**TRANSFAC (P-Match):** E2F **ConSite :** E2F	316 (+)	468 (+) Seq found: **TTTCCGAG**	 **TTTCGcgg**
PAPOLA	**TRANSFAC (P-Match):** E2F **ConSite :** E2F	188 (-)	No match	 **gcgGGAAA**
Total hits: 8/10 (80%)				
For TF *Relb*				
Genes	**Factor Name**	**Position in Human sequence (strand)**	**Position in Mouse sequence (strand)**	**Consensus sequence**
GYS1	**TRANSFAC (P-Match):** c-Rel **ConSite :** c-Rel	381 (-)	179 (-) Seq found: **GGAAACT**	 **GGAAAtcccc**
OGDH	**TRANSFAC (P-Match):** c-Rel **ConSite :** c-Rel	36 (+)	434 (+) Seq found: **CTGAAACC**	**caagaAAACC**
SLC25A4	**TRANSFAC (P-Match):** c-Rel **ConSite :** c-Rel	464 (-)	397 (-) Seq found: **gagCTTCC**	**tggagCTTCC**
STAR	**TRANSFAC (P-Match):** c-Rel	370 (+) 293 (-)	21 (+) Seq found:	**cttgaAAACC**

	ConSite : c-Rel		**AAACC**	
Ruvbl1	**TRANSFAC (P-Match):** c-Rel **ConSite :** c-Rel	53 (+)	No match	 **gggccTTTCC**
Sqrdl	**TRANSFAC (P-Match):** c-Rel **ConSite :** c-Rel	399 (+)	311 (+) Seq found: **CTTTC**	 **gggacCTTTC**
Abcd3	**TRANSFAC (P-Match):** c-Rel **ConSite :** c-Rel	1 (+)	54 (+) Seq found: **tctTTTCC**	 **tgtgtTTTCC**
Ngfrap1	No hits		c-Rel	
Map4k5	**TRANSFAC (P-Match):** c-Rel **ConSite :** c-Rel	93 (-)	142 (-) Seq found: **GGTTTTCAAA**	**GGTTTgcaaa**
Nfia	**TRANSFAC (P-Match):** c-Rel **ConSite :** c-Rel	621 (+)	No match	**cctccAAACC**
S100A1	**TRANSFAC (P-Match):** No hit; **ConSite :** c-Rel			
Actr3	**TRANSFAC (P-Match):** c-Rel **ConSite :** c-Rel	345 (-)	333 (-) Seq found: **GGAAA**	 **GGAAAacccg**
Ap3b1	**TRANSFAC (P-Match):** c-Rel **ConSite :** c-Rel	178 (+)	No match	 **tggcaTTTCC**

Btg1	**TRANSFAC (P-Match):** c-Rel **ConSite :** c-Rel	203 (+)	187 (+)	 **cggagCTTTC**
Total hits: 12/14 (85.71%)				

For TF *Sp3*

Genes	**Factor Name**	**Position in Human sequence (strand)**	**Position in Mouse sequence (strand)**	**Consensus sequence**
CLPTM1	**TRANSFAC (P-Match):** YY1 **ConSite :** Yin-yang	506 (-)	507 (-) Seq found: **ggaagATGGCggcgg**	**gagcgggaagATGGCggcgg**
PPARBP	**TRANSFAC (P-Match):** YY1 **ConSite :** No Match	497 (-)	485 (-)	**gttggggaagATGGCggcgg**
ARPC2	**TRANSFAC (P-Match):** YY1 **ConSite :** No Match	432 (-)	443 (-)	**gaagcggaaATGGCgccgc**
AP1M1	**TRANSFAC (P-Match):** No hit; **ConSite :** Yin-yang			
CIT	**TRANSFAC (P-Match):** No hit; **ConSite :** Yin-yang			
GABPA	**TRANSFAC (P-Match):** YY1 **ConSite :** No Match	49 (+) 212 (+)	430 (+) Seq found: **ctccGCCATcttt**	 **tacccGCCATcgcaatgcat** **cctccGCCATcttttcttcg**
Total hits: 4/6 (66.67%)				

For TF *STAT4*

Genes	**Factor Name**	**Position in Human sequence (strand)**	**Position in Mouse sequence (strand)**	**Consensus sequence**
ARPC2	**No hits**			
TMSB10	**TRANSFAC (P-Match):** STATx **ConSite :** STATx	235 (+)	213 (+) Seq found: **TTCCCg**	**TTCCCggaa**
Total hits: 1/2 (50%)				

We screened out common genes that are co-expressed with these TFs. In order to quantify the conservation of regulatory elements along these gene sequences, the core promoter regions (see Methods) were fed to P-Match [55] using all three available options for handling false discoveries. Basically, the output with option “minimizing false negative” considers merely minimal number of base pairs match and calls it a hit. Thus it improves its recall numbers (maximize loose-bound relevance at the cost of precision), with a huge list of *cis*-element candidates. We expect the false-positive rate to be extremely high for the predictions to be meaningful. Therefore, we did not discard this option. Among the other options, "minimize false positive" tries to find exact (~100%) PWM match and accounts for the most precise TF hits. The other option "minimize sum of both error rates" seems to take advantage from the best of both worlds (keeping balance on both recall and precision) and evens out high false discovery rates. To ensure better quality of our analysis, we considered only the option "minimize false positive", which maximizes the precision values without compromising too much with recall values. We summarize the sample results (precision >= 50%) for both human and mouse genomes in Table 3. The results for consulting ConSite are furnished as well. The consensus sequences (Logo-plots) for respective TFBS were extracted from TFM-Explorer [17].

Among eight TFs, our prediction performances for four TFs are encouraging while there is no highly correlated human pair showed up for the other four TFs. For instance, out of the 10 human genes that are associated with *E2F5* (E2F transcription factor 5), a member of *E2F* TF family, 8 genes (80% hit rate) carry the supposed binding sites. Comparing the sequence patterns of binding sites, we can say that almost all of them share the consensus **'TTTSSCGC'** where S could be a C or G. Out of these 8 human genes that are hit, 7 of the mouse orthologs also have the consensus sequence **'TTTSSCGC'** in the promoter region.

As per validation, we reviewed biological literature and found that there is a strong association between *E2F5* and *EZH2* (enhancer of zeste homolog 2), which appears in our result for binding site predictions. The *E2F* transcription factors, considered to be oncogenes [59], are key regulators of cell cycle progression [60], apoptosis and DNA damage response [61]. Loss of *E2F* functionality results in acute developmental effects [62]. It is also evidenced that *pRB-E2F* pathway (The retinoblastoma protein-*E2F*) strongly regulates the expressions of Polycomb group genes (PcG), one of which is *EZH2* [63, 64]. Again, *EZH2* is downstream of the *pRB–E2F* pathway and it is up-regulated during the cell proliferation and down-regulated during the cell differentiation phases [65].

Likewise, for the 14 human genes correlated with TF *RELB*, we have found 12 genes have the consensus sequence for *RELB* binding which achieve a hit rate of (85.71%). Among these 12 human genes, the mouse orthologs of 10 genes also contain some consensus sequence for *RELB* binding. Here, we found "**TTTCC"** as sense (+), or "**GGAAA"** as anti-sense (-) complementary, to be common motif with a number of out of pattern nucleotides around. Similarly, the precision rate for *sp3* (Sp3 transcription factor) is 66.67% (4 hits out of 6 genes) and for *STAT4* (signal transducer and activator of transcription 4) is 50% (1 hit from 2 genes). For *sp3*, listed genes have *cis*-elements containing motif "**GCCAT"** as sense (+), or "**ATGGC"** as anti-sense (-) counterpart. We believe that the hit rates for the rest are trivial due to the lack of sufficient correlated genes showing up.

3.2 Analyses of oncogenes

We selected a number of proto-oncogenes that are linked to various types of leukemia (AML, ALL, CML etc.) and breast cancer to illustrate our analyses. As discussed below, some correlated TF-gene pairs have been shown to bear regulatory relationship (Table 4 and Table 5). Moreover, some prospective candidates in our results actually comply with supporting biological literature reviews.

3.2.1 Leukemia genes

The *BCR* (Breakpoint cluster region) contains the Chromosome 22 breakpoint for the translocation that produces Philadelphia Chromosome [66], which is very often found in patients with CML (Chronic Myelogenous Leukemia). The translocation occurs between BCR (in Chromosome 22) and *ABL* (*v-abl* Abelson murine leukemia viral oncogene homolog in Chromosome 9q34), and this reciprocal translocation creates a tyrosine kinase, which is self-activated. Eventually this complex speeds up cell division, inhibits DNA repair, and causes genomic instability and blast effect in CML. Our analysis shows that *BCR* is up-regulated with *NR2F6* and *PCOLN3*, which complements previous studies [67, 68]. Our other candidates are *CENPB* (Centromere protein B) and *NCOR2* (Nuclear receptor co-repressor 2) as positively correlated transcription factors and *SP3* (Sp3 transcription factor) as negatively correlated transcription factors.

One of the well-recognized oncogenes for AML (Acute Myelogenous Leukemia) is *MLL* (myeloid/lymphoid or mixed-lineage leukemia) [69-72]. In our analysis, we found *MDM2* (transformed 3T3 cell double minute 2, p53 binding protein), which is a potential target of tumor suppressor protein p53, to be up-regulated with *MLL*. Over-expression of *MDM2* may cause excessive inactivation of tumor protein p53, deteriorating its tumor suppressor function [73, 74].

The *DEK* (DEK oncogene) gene produces a fusion with the *CAN* protein in a subtype of AML patients [75]. We found *HMGB1* (high-mobility group box 1) in our analysis, and it was also reported in the literature as a gene which controls the binding behavior of DEK [76]. Another transcription factor correlated with DEK is *sp3* (Sp3 transcription factor) [77].

We also furnished correlated transcription factors for *LMO2* (LIM domain only 2), *ETV6* (ets variant gene 6), *RUNX1* (runt-related transcription factor 1, also called *AML1*), *BTG1* (B-cell translocation gene 1, anti-proliferative) *FCGR2B* (Fc fragment of IgG, low affinity IIb, receptor, Loc. 1q23), *LCK* (lymphocyte-specific protein tyrosine kinase, Loc. 1p34.3), which are relevant to leukemia.

Table 4: The list of transcription factors highly correlated to leukemia genes

Gene	Transcription Factor	Correlation Coefficient	Gene	Transcription Factor	Correlation Coefficient
	ZMIZ2	0.6532		*PLEK*	0.8295
	DPP7	0.6446		*SP110*	0.7726
	EPN1	0.5957		*DOCK2*	0.7502
	GRINA	0.5837		*TMSB4Y*	0.7468
BCR	*HMOX2*	0.5641	*BTG1*	*CD53*	0.7301
	GABPA	-0.6634		*GCSH*	-0.5434
	PPARBP	-0.6360		*BMPR1A*	-0.5015
	TFAM	-0.6298		*LRP6*	-0.4736
	MAPK8IP3	-0.6247		*PFN2*	-0.4682
	STAG2	-0.6236		*ZDHHC7*	-0.4587
	ADAMTS1	0.6352		*RAC2*	0.9286
	YWHAZ	0.6003		*TRB@*	0.9107
	IFI16	0.5976		*LIMD2*	0.9086
	KLF4	0.5855		*CD3D*	0.8960
MYC	*TUBA3*	0.5813	*LCK*	*DOCK2*	0.8875
	MYO10	-0.5300		*NCKAP1*	-0.6768
	GCSH	-0.5190		*LRP6*	-0.6414
	XPO4	-0.5034		*RBM9*	-0.6216
	DDAH1	-0.4945		*YAP1*	-0.6174

	MAPRE3	-0.4925		*BMPR1A*	-0.6041
	HMGB1	0.7934		*FCGR2C*	0.8553
	STAG2	0.7594		*FCER1G*	0.8087
	GDI2	0.7302		*AIF1*	0.7920
	PSMA4	0.7113		*HCLS1*	0.7877
DEK	*PPARBP*	0.7073	*FCGR2B*	*HEM1*	0.7851
	WWP2	-0.5237		*TOM1L1*	-0.5286
	ATP6V0A1	-0.5220		*MTX2*	-0.4815
	QDPR	-0.5002		*PFN2*	-0.4796
	EPN1	-0.4984		*HDLBP*	-0.4759
	GRINA	-0.4897		*H1F0*	-0.4735
	DNCH1	0.6081		*SSH2*	0.5702
	AKAP11	0.5910		*MSL3L1*	0.5087
	MPHOSPH1	0.5363		*MAP4K4*	0.5024
	DNCH1	0.5282		*BRD4*	0.4790
MLL	*TERF2*	0.4534	*RUNX1*	*GMFG*	0.4715
	PRDX6	-0.4987		*CRYL1*	-0.5235
	KIAA0152	-0.4449		*FLJ20315*	-0.4564
	IL1R1	-0.4410		*GJA7*	-0.4508
	TM9SF1	-0.4235		*AIG1*	-0.4493
	MID1	-0.4225		*CRYZ*	-0.4377
	PTH	0.5959		*INPP4B*	0.6494
	GATA3	0.5737		*MAP3K8*	0.6462
	SYK	0.5693		*CHDH*	0.6392
	PVALB	0.5657		*SWAP70*	0.6307
ETV6	*PTN*	0.5468	*LMO2*	*PTTG1*	0.6187
	KIAA0934	-0.4292		*NBEA*	-0.4997
	PMP22	-0.4246		*CDC42BPA*	-0.4722
	FBN1	-0.4151		*MXI1*	-0.4624
	TCF8	-0.4138		*XK*	-0.4319
	NT5E	-0.4064		*EXTL2*	-0.4251

Our analyses on oncogenes provide some supporting evidence for the correlations between transcription factors and corresponding genes. Most often, the findings are consistent with what we found in biological literature surveys. Also we find a number of significant candidates that may be quite relevant to cancer studies. For example, *NR2F6*, which is a nuclear receptor, is positively correlated with breast cancer gene (*ErbB2*) and one of the leukemia genes (*BCR*), while it is inversely correlated with other leukemia gene (*DEK*). Collectively, we may analyze the inter-relationships among leukemia genes and correlated transcription factors (see Table 6). Similarly, we discovered identical expression pattern among *BTG1, FCGR2B* and *LCK* genes (see Table 7). In fact, *FCGR2B* and *LCK* show up in the B-Cell Receptor Signaling Pathway and the T-Cell Receptor Signaling Pathway respectively, which seems to be the reason for this concurrence.

Table 5: The list of transcription factors correlated to breast cancer genes

Gene	Transcription Factor	Correlation Coefficient
	FOXM1	0.4913
	HMGA1	0.5195
	HHEX	0.4388
	MCM5	0.4048
	GMEB1	0.395

BRCA1	*PTTG1*	0.3839
	E2F1	0.3816
	NFIB	-0.4069
	FOSB	-0.2364
	EGR1	-0.2075
	TBX3	-0.1882
	LPIN1	0.7520
	SPINT1	0.6570
	MKNK2	0.6365
	DDR1	0.6153
	NR2F6	0.4288
	MEF2C	-0.5445
ErbB2	*NIN*	-0.52263
	MBNL1	-0.52012
	SSBP2	-0.51904
	ZFP91	-0.4626

3.2.2 Breast cancer genes

Unlike leukemia oncogenes, we applied our analysis on two well-known breast cancer genes, namely *BRCA1* and *ErbB2*. Because of small number of genes, clustering is not feasible to discover the interactions. However, we found some associations among these genes, which are backed by literature.

Certain mutations of *BRCA1* (Breast Cancer 1, early onset) cause approximately 40% of inherited breast cancers and more than 80% of inherited breast and ovarian cancers [78-89]. Our analysis reveals that *FOXM1, E2F1, PTTG1, HMGA1, GMEB1, MCM5* and *HHEX* transaction factor genes are positively correlated to the expression of *BRCA1*, while expressions of *NFIB, EGR1, FOSB,* and *TBX3* appear to be inversely correlated to that of *BRCA1*.

Table 6: Association analysis of *BCR,* and *DEK* with transcription factors in terms of regulation pattern

TF	*BCR*	*DEK*
NR2F6	+0.4302	-0.3505
NCOR2	+0.4343	-0.3916
ZNF74	+0.4504	-0.2813
PCOLN3	+0.3442	-0.2912
NR1D1	+0.3491	-0.2866
HMX1	+0.4477	-0.4396
HMGB1	-0.5228	+0.7934
ATF1	-0.5145	+0.5373
SP3	-0.5664	+0.6177
TFAM	-0.6298	+0.6165
GABPA	-0.6634	+0.6907
NPM1	-0.6390	+0.6731
ELF1	-0.4434	+0.6473

FOXM1 (Forkhead box M1) has been found to be responsible for epithelial ovarian tumors correlated with malignancy [90] and it is over-expressed in transcriptional regulations. There is much evidence that *E2F1* (E2F transcription factor 1) is related to control of *BRCA1*.

Hennighausen [91] simulated the mouse model of breast cancer and established the presence of *E2F1* in cancer-causing. Again, the *E2F* family plays a crucial role in the control of cell cycle and action of tumor suppressor proteins and is also a target of the transforming proteins of small DNA tumor viruses. So, *E2F*-responsive promoters appear to be more active in tumor cells relative to normal cells. Recently, pharmacogenomics studies showed that murine-derived anticancer agents, namely YondelisTM (Trabectedin, ET-743) [92], target markers like *PTTG1* (pituitary tumor-transforming 1) and *HHEX* (Homeobox, hematopoietically expressed). Baldassarre et. al. [93] found strong interactions between *HMGA1* (high mobility group AT-hook 1) and *BRCA1* in sporadic breast carcinoma. In fact, transcriptional binding sites of *BRCA1* are edged upstream and downstream by AT-rich sequences which represent preferred binding sites for *HMGA1*. Moreover, Wan et. al. [67] reported the association of *GMEB1* (Glucocorticoid modulatory element binding protein 1 and *MCM5* (Minichromosome maintenance deficient 5, cell division cycle 46). In negative counterpart, that *EGR1* (Early growth response 1) inversely act towards the *BRCA1* expression is evidenced by Cooper [94] and Robson [95]. Substantial facts of *TBX3* (T-box 3,) being correlated with *BRCA1* has been also reported [96].

Table 7: Association analysis of *BTG1, FCGR2B* and *LCK* with transcription factors in terms of regulation pattern.

TF	*BTG1*	*FCGR2B*	*LCK*
ETS1	+0.7230	+0.6618	+0.8468
SP110	+0.7726	+0.6827	+0.7928
SP140	+0.6696	+0.5598	+0.7821
HCLS1	+0.7297	+0.7877	+0.8101
TNFAIP3	+0.5729	+0.6513	+0.5316
IFI16	+0.5869	+0.6094	+0.5391
TRIM22	+0.5860	+0.6056	+0.6201
ARNT2	-0.6084	-0.2948	-0.2755
ARNTL	+0.6701	+0.7616	+0.5599
ELF1	+0.6083	+0.5397	+0.7234
RBM9	-0.3579	-0.3460	-0.6216

ErbB2/HER2 (HER2/neu, v-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/ glioblastoma derived oncogene homolog) [97-101] is a proto-oncogene located at 17q11.2-q12, 17q21.1 and belongs to the epithelial growth factor receptor (EGFR). It codes for a cell membrane surface-bound receptor tyrosine kinase that is expressed abundantly by transcriptional regulation in approximately 20%-40% of breast cancer and it is notably treated as a target gene for therapy. Often it plays a vital role in cell growth, differentiation and metastasis. We found several TFs, like *NR2F6, RORC, TEAD3, IRX3, SOX13, PCOLN3*, as positively expressed, and *MEF2C, ZFP91, MEF2C, ZFP1* as negatively correlated with *ErbB2*. *NR2F6* (Nuclear receptor subfamily 2, group F, member 6) was commented to be coupled with the expression of *HER2* [68]. Also in-vitro studies showed that he phosphorylation of the exogenous *MEF2C* (Myocyte enhancer factor 2C) is dramatically increased by epithelial growth factor (EGF) and thus it refers to its association with controls of *ErbB2* expression [102].

4 Conclusions

By analyzing tissue-specific gene expression profiles of human and mouse, we produced a list of putative *trans*-elements and their associated genes. We demonstrated that the use of decay function instead of linear function as regressor is more appropriate for capturing possible regulatory relationship. As a measure of the quality of the predicted *trans*-elements, we focus on the set of transcription factors and genes that share the same gene names in both mouse and human data sets. Our analysis of oncogenes provides further assurance of the quality of the predicted *trans*-elements. Note that a highly correlated TF-gene pair that is not currently known to bear regulatory relationship may still be a correct *trans*-element prediction that is yet to be validated. ChIP-chip experiment is designed to test protein-DNA binding, where the binding of a transcription factor to a promoter region of a gene provides necessary but not sufficient evidence for transcription regulation. As a possible further step to confirm the regulatory relationship, the TF-gene pairs and their correlation coefficients constructed here may serve as a source of reference for additional confirmation of ChIP-chip results.

5 Acknowledgements

This research is supported partially by NSF CCR-0208935, NIH NCRR Research Center at Minority Institutions Grant number 2G12RR1364-06A1 and funding from the University of Texas System. We are also thankful to Dr. Stephen Harris and Dr. Carolina Livi for their valuable comments and suggestions.

References

[1] J. D. McPherson, M. Marra, L. Hillier, R. H. Waterston, A. Chinwalla, J. Wallis, et al., *A physical map of the human genome*, Nature, 409, pp. 934-41, 2001.

[2] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, et al., *Initial sequencing and analysis of the human genome*, Nature, 409, pp. 860-921, 2001.

[3] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, et al., *The sequence of the human genome*, Science, 291, pp. 1304-51, 2001.

[4] J. W. Fickett and W. W. Wasserman, *Discovery and modeling of transcriptional regulatory regions*, Curr Opin Biotechnol, 11, pp. 19-24, 2000.

[5] G. Pavesi, G. Mauri and G. Pesole, *In silico representation and discovery of transcription factor binding sites*, Brief Bioinform, 5, pp. 217-36, 2004.

[6] L. A. McCue, W. Thompson, C. S. Carmack and C. E. Lawrence, *Factors influencing the identification of transcription factor binding sites by cross-species comparison*, Genome Res, 12, pp. 1523-32, 2002.

[7] R. Osada, E. Zaslavsky and M. Singh, *Comparative analysis of methods for representing and searching for transcription factor binding sites*, Bioinformatics, 20, pp. 3516-25, 2004.

[8] S. Sinha and M. Tompa, *Discovery of novel transcription factor binding sites by statistical overrepresentation*, Nucleic Acids Res, 30, pp. 5549-60, 2002.

[9] S. Sinha and M. Tompa, *YMF: A program for discovery of novel transcription factor binding sites by statistical overrepresentation*, Nucleic Acids Res, 31, pp. 3586-8, 2003.

[10] A. M. Moses, D. Y. Chiang, D. A. Pollard, V. N. Iyer and M. B. Eisen, *MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model*, Genome Biol, 5, pp. R98, 2004.

[11] S. Aerts, G. Thijs, B. Coessens, M. Staes, Y. Moreau and B. De Moor, *Toucan: deciphering the cis-regulatory logic of coregulated genes*, Nucleic Acids Res, 31, pp. 1753-64, 2003.

[12] G. Z. Hertz and G. D. Stormo, *Identifying DNA and protein patterns with statistically significant alignments of multiple sequences*, Bioinformatics, 15, pp. 563-77, 1999.

[13] G. D. Stormo, *DNA binding sites: representation and discovery*, Bioinformatics, 16, pp. 16-23, 2000.

[14] W. B. Alkema, O. Johansson, J. Lagergren and W. W. Wasserman, *MSCAN: identification of functional clusters of transcription factor binding sites*, Nucleic Acids Res, 32, pp. W195-8, 2004.

[15] O. Johansson, W. Alkema, W. W. Wasserman and J. Lagergren, *Identification of functional clusters of transcription factor binding motifs in genome sequences: the MSCAN algorithm*, Bioinformatics, 19 Suppl 1, pp. i169-76, 2003.

[16] J. Zheng, J. Wu and Z. Sun, *An approach to identify over-represented cis-elements in related sequences*, Nucleic Acids Res, 31, pp. 1995-2005, 2003.

[17] M. Defrance and H. Touzet, *Predicting transcription factor binding sites using local over-representation and comparative genomics*, BMC Bioinformatics, 7, pp. 396, 2006.

[18] P. E. Boardman, S. G. Oliver and S. J. Hubbard, *SiteSeer: Visualisation and analysis of transcription factor binding sites in nucleotide sequences*, Nucleic Acids Res, 31, pp. 3572-5, 2003.

[19] A. E. Kel, E. Gossling, I. Reuter, E. Cheremushkin, O. V. Kel-Margoulis and E. Wingender, *MATCH: A tool for searching transcription factor binding sites in DNA sequences*, Nucleic Acids Res, 31, pp. 3576-9, 2003.

[20] M. C. Frith, M. C. Li and Z. Weng, *Cluster-Buster: Finding dense clusters of motifs in DNA sequences*, Nucleic Acids Res, 31, pp. 3666-8, 2003.

[21] N. Rajewsky, M. Vergassola, U. Gaul and E. D. Siggia, *Computational detection of genomic cis-regulatory modules applied to body patterning in the early Drosophila embryo*, BMC Bioinformatics, 3, pp. 30, 2002.

[22] T. Werner, *Models for prediction and recognition of eukaryotic promoters*, Mamm Genome, 10, pp. 168-75, 1999.

[23] H. J. Bussemaker, H. Li and E. D. Siggia, *Regulatory element detection using correlation with expression*, Nat Genet, 27, pp. 167-71, 2001.

[24] E. M. Conlon, X. S. Liu, J. D. Lieb and J. S. Liu, *Integrating regulatory motif discovery and genome-wide expression analysis*, Proc Natl Acad Sci U S A, 100, pp. 3339-44, 2003.

[25] C. T. Workman and G. D. Stormo, *ANN-Spec: a method for discovering transcription factor binding sites with improved specificity*, Pac Symp Biocomput, pp. 467-78, 2000.

[26] M. C. Frith, U. Hansen, J. L. Spouge and Z. Weng, *Finding functional sequence elements by multiple local alignment*, Nucleic Acids Res, 32, pp. 189-200, 2004.

[27] A. V. Favorov, M. S. Gelfand, A. V. Gerasimova, D. A. Ravcheev, A. A. Mironov and V. J. Makeev, *A Gibbs sampler for identification of symmetrically structured, spaced*

DNA motifs with improved estimation of the signal length, Bioinformatics, 21, pp. 2240-5, 2005.

[28] K. Ellrott, C. Yang, F. M. Sladek and T. Jiang, *Identifying transcription factor binding sites through Markov chain optimization*, Bioinformatics, 18 Suppl 2, pp. S100-9, 2002.

[29] W. Ao, J. Gaudet, W. J. Kent, S. Muttumu and S. E. Mango, *Environmentally induced foregut remodeling by PHA-4/FoxA and DAF-12/NHR*, Science, 305, pp. 1743-6, 2004.

[30] K. A. Frazer, L. Elnitski, D. M. Church, I. Dubchak and R. C. Hardison, *Cross-species sequence comparisons: a review of methods and available resources*, Genome Res, 13, pp. 1-12, 2003.

[31] S. Batzoglou, L. Pachter, J. P. Mesirov, B. Berger and E. S. Lander, *Human and mouse gene structure: comparative analysis and application to exon prediction*, Genome Res, 10, pp. 950-8, 2000.

[32] N. Jareborg and R. Durbin, *Alfresco--a workbench for comparative genomic sequence analysis*, Genome Res, 10, pp. 1148-57, 2000.

[33] A. Ureta-Vidal, L. Ettwiller and E. Birney, *Comparative genomics: genome-wide analysis in metazoan eukaryotes*, Nat Rev Genet, 4, pp. 251-62, 2003.

[34] B. Lenhard, A. Sandelin, L. Mendoza, P. Engstrom, N. Jareborg and W. W. Wasserman, *Identification of conserved regulatory elements by comparative genome analysis*, J Biol, 2, pp. 13, 2003.

[35] B. Lenhard, C. Wahlestedt and W. W. Wasserman, *GeneLynx mouse: integrated portal to the mouse genome*, Genome Res, 13, pp. 1501-4, 2003.

[36] S. R. Eddy, *A model of the statistical power of comparative genome sequence analysis*, PLoS Biol, 3, pp. e10, 2005.

[37] L. Duret and P. Bucher, *Searching for regulatory elements in human noncoding sequences*, Curr Opin Struct Biol, 7, pp. 399-406, 1997.

[38] V. X. Jin, A. Rabinovich, S. L. Squazzo, R. Green and P. J. Farnham, *A computational genomics approach to identify cis-regulatory modules from chromatin immunoprecipitation microarray data--A case study using E2F1*, Genome Res, 2006.

[39] V. X. Jin, G. A. Singer, F. J. Agosto-Perez, S. Liyanarachchi and R. V. Davuluri, *Genome-wide analysis of core promoter elements from conserved human and mouse orthologous pairs*, BMC Bioinformatics, 7, pp. 114, 2006.

[40] S. Y. Kim and Y. Kim, *Genome-wide prediction of transcriptional regulatory elements of human promoters using gene expression and promoter analysis data*, BMC Bioinformatics, 7, pp. 330, 2006.

[41] K. Lindblad-Toh, E. S. Lander, J. D. McPherson, R. H. Waterston, J. Rodgers and E. Birney, *Progress in sequencing the mouse genome*, Genesis, 31, pp. 137-41, 2001.

[42] J. Kawai, A. Shinagawa, K. Shibata, M. Yoshino, M. Itoh, Y. Ishii, et al., *Functional annotation of a full-length mouse cDNA collection*, Nature, 409, pp. 685-90, 2001.

[43] Y. Okazaki, M. Furuno, T. Kasukawa, J. Adachi, H. Bono, S. Kondo, et al., *Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs*, Nature, 420, pp. 563-73, 2002.

[44] R. C. Hardison, J. Oeltjen and W. Miller, *Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome*, Genome Res, 7, pp. 959-66, 1997.

[45] R. D. Emes, L. Goodstadt, E. E. Winter and C. P. Ponting, *Comparison of the genomes of human and mouse lays the foundation of genome zoology*, Hum Mol Genet, 12, pp. 701-9, 2003.

[46] W. W. Wasserman, M. Palumbo, W. Thompson, J. W. Fickett and C. E. Lawrence, *Human-mouse genome comparisons to locate regulatory sites*, Nat Genet, 26, pp. 225-8, 2000.

[47] E. T. Dermitzakis, A. Reymond, R. Lyle, N. Scamuffa, C. Ucla, S. Deutsch, et al., *Numerous potentially functional but non-genic conserved sequences on human chromosome 21*, Nature, 420, pp. 578-82, 2002.

[48] R. Shyamsundar, Y. H. Kim, J. P. Higgins, K. Montgomery, M. Jorden, A. Sethuraman, et al., *A DNA microarray survey of gene expression in normal human tissues*, Genome Biol, 6, pp. R22, 2005.

[49] W. Zhang, Q. D. Morris, R. Chang, O. Shai, M. A. Bakowski, N. Mitsakakis, et al., *The functional landscape of mouse gene expression*, J Biol, 3, pp. 21, 2004.

[50] *Entrez Gene* http://www.ncbi.nlm.nih.gov/entrez/http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=gene,

[51] *RefSeq* www.ncbi.nlm.nih.gov/RefSeq/,

[52] W. W. Wasserman and A. Sandelin, *Applied bioinformatics for the identification of regulatory elements*, Nat Rev Genet, 5, pp. 276-87, 2004.

[53] V. Matys, E. Fricke, R. Geffers, E. Gossling, M. Haubrock, R. Hehl, et al., *TRANSFAC: transcriptional regulation, from patterns to profiles*, Nucleic Acids Res, 31, pp. 374-8, 2003.

[54] *TRANSFAC-The Transcription Factor Database* http://www.gene-regulation.com/pub/databases.html,

[55] D. S. Chekmenev, C. Haid and A. E. Kel, *P-Match: transcription factor binding site search by combining patterns and weight matrices*, Nucleic Acids Res, 33, pp. W432-7, 2005.

[56] A. Sandelin, W. W. Wasserman and B. Lenhard, *ConSite: web-based prediction of regulatory elements using cross-species comparison*, Nucleic Acids Res, 32, pp. W249-52, 2004.

[57] *Supplementary data link for this page* http://www.cs.utsa.edu/~hugelab/supDataTF.html.,

[58] C. Kunsch, S. M. Ruben and C. A. Rosen, *Selection of optimal kappa B/Rel DNA-binding motifs: interaction of both subunits of NF-kappa B with DNA is required for transcriptional activation*, Mol Cell Biol, 12, pp. 4412-21, 1992.

[59] P. K. Tsantoulis and V. G. Gorgoulis, *Involvement of E2F transcription factor family in cancer*, Eur J Cancer, 41, pp. 2403-14, 2005.

[60] C. Attwooll, E. Lazzerini Denchi and K. Helin, *The E2F family: specific functions and overlapping interests*, Embo J, 23, pp. 4709-16, 2004.

[61] H. Muller, A. P. Bracken, R. Vernell, M. C. Moroni, F. Christians, E. Grassilli, et al., *E2Fs regulate the expression of genes involved in differentiation, development, proliferation, and apoptosis*, Genes Dev, 15, pp. 267-85, 2001.

[62] G. J. Lindeman, L. Dagnino, S. Gaubatz, Y. Xu, R. T. Bronson, H. B. Warren, et al., *A specific, nonproliferative role for E2F-5 in choroid plexus function revealed by gene targeting*, Genes Dev, 12, pp. 1092-8, 1998.

[63] B. Czermin, R. Melfi, D. McCabe, V. Seitz, A. Imhof and V. Pirrotta, *Drosophila enhancer of Zeste/ESC complexes have a histone H3 methyltransferase activity that marks chromosomal Polycomb sites*, Cell, 111, pp. 185-96, 2002.

[64] A. Kuzmichev, K. Nishioka, H. Erdjument-Bromage, P. Tempst and D. Reinberg, *Histone methyltransferase activity associated with a human multiprotein complex containing the Enhancer of Zeste protein*, Genes Dev, 16, pp. 2893-905, 2002.

[65] A. P. Bracken, D. Pasini, M. Capra, E. Prosperini, E. Colli and K. Helin, *EZH2 is downstream of the pRB-E2F pathway, essential for proliferation and amplified in cancer*, Embo J, 22, pp. 5323-35, 2003.

[66] R. Kurzrock, H. M. Kantarjian, B. J. Druker and M. Talpaz, *Philadelphia chromosome-positive leukemias: from basic mechanisms to molecular therapeutics*, Ann Intern Med, 138, pp. 819-30, 2003.

[67] D. Wan, Y. Gong, W. Qin, P. Zhang, J. Li, L. Wei, et al., *Large-scale cDNA transfection screening for genes related to cancer development and progression*, Proc Natl Acad Sci U S A, 101, pp. 15724-9, 2004.

[68] M. A. Gieseg, T. Cody, M. Z. Man, S. J. Madore, M. A. Rubin and E. P. Kaldjian, *Expression profiling of human renal carcinomas with functional taxonomic analysis*, BMC Bioinformatics, 3, pp. 26, 2002.

[69] X. S. Puente, G. Velasco, A. Gutierrez-Fernandez, J. Bertranpetit, M. C. King and C. Lopez-Otin, *Comparative analysis of cancer genes in the human and chimpanzee genomes*, BMC Genomics, 7, pp. 15, 2006.

[70] M. Stanulla, H. J. Schunemann, S. Thandla, M. L. Brecher and P. D. Aplan, *Pseudo-rearrangement of the MLL gene at chromosome 11q23: a cautionary note on genotype analysis of leukaemia patients*, Mol Pathol, 51, pp. 85-9, 1998.

[71] L. W. Deng, I. Chiu and J. L. Strominger, *MLL 5 protein forms intranuclear foci, and overexpression inhibits cell cycle progression*, Proc Natl Acad Sci U S A, 101, pp. 757-62, 2004.

[72] S. Yamamoto, M. Nishi, K. Taniguchi, M. Imayoshi, Y. Ogata, M. Iwanaga, et al., *Partial tandem duplication of MLL gene in acute myeloid leukemia with translocation (11;17)(q23;q12-21)*, Am J Hematol, 80, pp. 46-9, 2005.

[73] J. L. Best, C. A. Amezcua, B. Mayr, L. Flechner, C. M. Murawsky, B. Emerson, et al., *Identification of small-molecule antagonists that inhibit an activator: coactivator interaction*, Proc Natl Acad Sci U S A, 101, pp. 17622-7, 2004.

[74] D. Wiederschain, H. Kawai, J. Gu, A. Shilatifard and Z. M. Yuan, *Molecular basis of p53 functional inactivation by the leukemic protein MLL-ELL*, Mol Cell Biol, 23, pp. 4230-46, 2003.

[75] T. Waldmann, I. Scholten, F. Kappes, H. G. Hu and R. Knippers, *The DEK protein--an abundant and ubiquitous constituent of mammalian chromatin*, Gene, 343, pp. 1-9, 2004.

[76] T. Waldmann, M. Baack, N. Richter and C. Gruss, *Structure-specific binding of the proto-oncogene protein DEK to DNA*, Nucleic Acids Res, 31, pp. 7003-10, 2003.

[77] D. Y. Lin, H. I. Fang, A. H. Ma, Y. S. Huang, Y. S. Pu, G. Jenster, et al., *Negative modulation of androgen receptor transcriptional activity by Daxx*, Mol Cell Biol, 24, pp. 10529-41, 2004.

[78] K. Metcalfe, H. T. Lynch, P. Ghadirian, N. Tung, I. Olivotto, E. Warner, et al., *Contralateral breast cancer in BRCA1 and BRCA2 mutation carriers*, J Clin Oncol, 22, pp. 2328-35, 2004.

[79] W. D. Foulkes, K. Metcalfe, P. Sun, W. M. Hanna, H. T. Lynch, P. Ghadirian, et al., *Estrogen receptor status in BRCA1- and BRCA2-related breast cancer: the influence of age, grade, and histological type*, Clin Cancer Res, 10, pp. 2029-34, 2004.

[80] A. Antoniou, P. D. Pharoah, S. Narod, H. A. Risch, J. E. Eyfjord, J. L. Hopper, et al., *Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case Series unselected for family history: a combined analysis of 22 studies*, Am J Hum Genet, 72, pp. 1117-30, 2003.

[81] G. L. Barnett and C. A. Friedrich, *Recent developments in ovarian cancer genetics*, Curr Opin Obstet Gynecol, 16, pp. 79-85, 2004.

[82] D. C. Daniel, *Highlight: BRCA1 and BRCA2 proteins in breast cancer*, Microsc Res Tech, 59, pp. 68-83, 2002.

[83] S. L. Ding, L. F. Sheu, J. C. Yu, T. L. Yang, B. F. Chen, F. J. Leu, et al., *Abnormality of the DNA double-strand-break checkpoint/repair genes, ATM, BRCA1 and TP53, in breast cancer is related to tumour grade*, Br J Cancer, 90, pp. 1995-2001, 2004.

[84] A. Liede, B. Y. Karlan and S. A. Narod, *Cancer risks for male carriers of germline mutations in BRCA1 or BRCA2: a review of the literature*, J Clin Oncol, 22, pp. 735-42, 2004.

[85] S. N. Powell and L. A. Kachnic, *Roles of BRCA1 and BRCA2 in homologous recombination, DNA replication fidelity and the cellular response to ionizing radiation*, Oncogene, 22, pp. 5784-91, 2003.

[86] R. Scully and N. Puget, *BRCA1 and BRCA2 in hereditary breast cancer*, Biochimie, 84, pp. 95-102, 2002.

[87] A. A. Tutt A, *The relationship between the roles of BRCA genes in DNA repair and cancer predisposition*, Trends Mol Med, 8, pp. 571-6, 2002.

[88] A. R. Venkitaraman, *Cancer susceptibility and the functions of BRCA1 and BRCA2*, Cell, 108, pp. 171-82, 2002.

[89] B. N. Nadeau G, Moisan A, Lemieux KM, Cayanan C, Monteiro AN, Gaudreau L, *BRCA1 can stimulate gene transcription by a unique mechanism. EMBO Rep. 2000 Sep 15; 1(3): 260-265.*, EMBO Rep., 1, pp. 260-265, 2000.

[90] P. S. Warrenfeltz S, Datta S, Kraemer ET, Benigno B, McDonald JF, *Gene expression profiling of epithelial ovarian tumors correlated with malignant potential.*, Mol Cancer, 3, 2004.

[91] L. Hennighausen, *Mouse models for breast cancer*, Breast Cancer Res, 2, pp. 2-7, 2000.

[92] J. Jimeno, M. Aracil and J. C. Tercero, *Adding pharmacogenomics to the development of new marine-derived anticancer agents*, J Transl Med, 4, pp. 3, 2006.

[93] B. S. Baldassarre G, Belletti B, Thakur S, Pentimalli F, Trapasso F, Fedele M, Pierantoni G, Croce CM, Fusco A., *Negative Regulation of BRCA1 Gene Expression by HMGA1 Proteins Accounts for the Reduced BRCA1 Protein Levels in Sporadic Breast Carcinoma.*, Mol Cell Biol., Apr; 23, pp. 2225-2238, 2003.

[94] C. S. Cooper, *Applications of microarray technology in breast cancer research*, Breast Cancer Res, 3, pp. 158-75, 2001.

[95] T. Robson and D. G. Hirst, *Transcriptional Targeting in Cancer Gene Therapy*, J Biomed Biotechnol, 2003, pp. 110-137, 2003.

[96] A. L. Welm, *AACR Special Conference: Advances in Breast Cancer Research--Genetics, Biology, and Clinical Implications, Huntington Beach, California, USA, 8-12 October 2003*, Breast Cancer Res, 6, pp. E6, 2004.

[97] J. Bange, E. Zwick and A. Ullrich, *Molecular targets for breast cancer therapy and prevention*, Nat Med, 7, pp. 548-52, 2001.

[98] T. Kute, C. M. Lack, M. Willingham, B. Bishwokama, H. Williams, K. Barrett, et al., *Development of Herceptin resistance in breast cancer cells*, Cytometry A, 57, pp. 86-93, 2004.

[99] S. Menard, S. M. Pupa, M. Campiglio and E. Tagliabue, *Biologic and therapeutic role of HER2 in cancer*, Oncogene, 22, pp. 6570-8, 2003.

[100] R. Nahta and F. J. Esteva, *HER-2-targeted therapy: lessons learned and future directions*, Clin Cancer Res, 9, pp. 5078-84, 2003.

[101] D. Yu and M. C. Hung, *Overexpression of ErbB2 in cancer and ErbB2-targeting strategies*, Oncogene, 19, pp. 6115-21, 2000.

[102] D.-R. E. Esparís-Ogando A, Montero JC, Yuste L, Crespo P, Pandiella A., *Erk5 Participates in Neuregulin Signal Transduction and Is Constitutively Active in Breast Cancer Cells Overexpressing ErbB2.*, Mol Cell Biol., Jan; 22, pp. 270-285., 2002.

Additional files

Additional file 1– Tf-gene.bmp

It contains the figure used in the manuscript

Additional file 2– Tf-gene1.bmp

It contains the figure used in the manuscript